Archiving in Minnesota: System Model and Implementation

Prepared for Workshop 131, TRB 2003

Dr. Taek Mu Kwon

Transportation Data Research Laboratory

Northland Advanced Transportation Systems Research Laboratories

Department of Electrical and Computer Engineering

Outline

- TDRL Introduction
- Centralization of Data
- Data hierarchies
- TDRL archive structure
- Conclusion

Transportation Data Research Laboratory (TDRL)

- TDRL is a part of NATSRL along with ASRL (Advanced Sensor Research Laboratory) at the University of Minnesota Duluth.
- Established to provide on-line ITS data resources for Minnesota
- TDRL focuses on research issues concerning large-scaled transportation data.

Sun
microsystems

S PROJECT
Neural Networks
Networks
Neural Networks

TDRL Data Model: Distributed Computing Model

- Centralization of data
- Decentralization (distribution) of computing

Advantages of Centralization of Data

- Efficient single point management
- Unified data format
- Consistent version control
- Large scale data warehousing by experts
- Efficient archiving and sharing
- Large scale data analysis

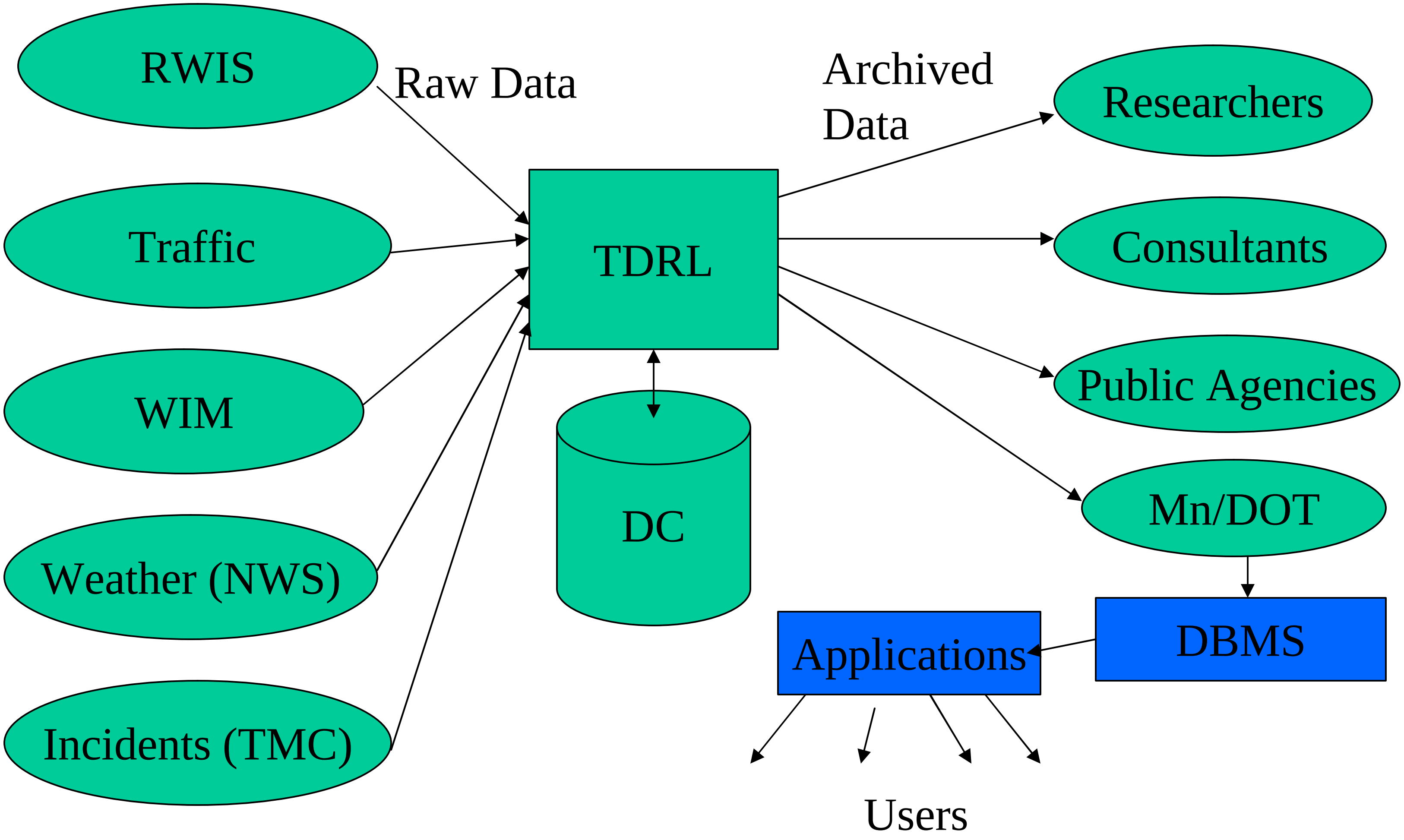
Advantages of Centralization of Data, Cont.

- Minimization of redundant efforts
- Efficient data integration and organization
- Secure archives (file backup, UPS, firewall)
- Easy cross reference and analysis
- Minimization of confusion in data
- Single point unified data quality control
- Advanced data application developments

Advantages of Centralization of Data, Cont.

- Easy and cheaper to upgrade computers
- Easy to maintain new versions of software
- Maximize data sharing among departments, public, and private sectors
- Easy to obtain user feedback and analyze
- Specialize data help
- Easy access of data

On-Line Data Source/Supply

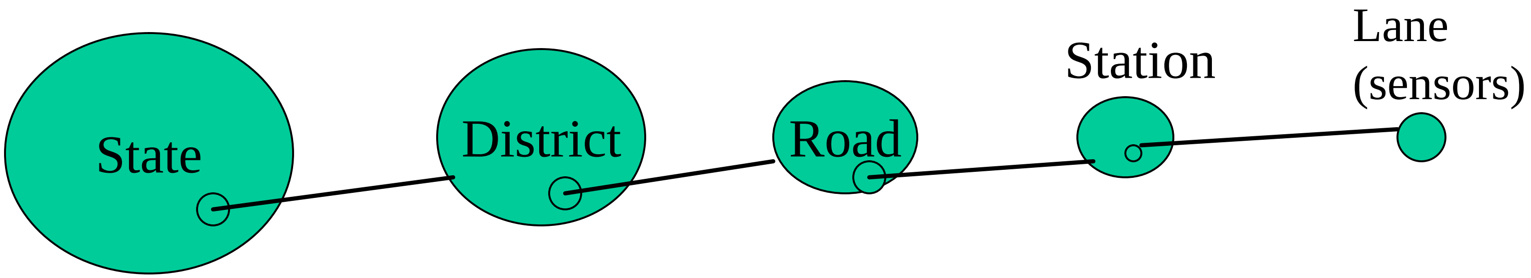


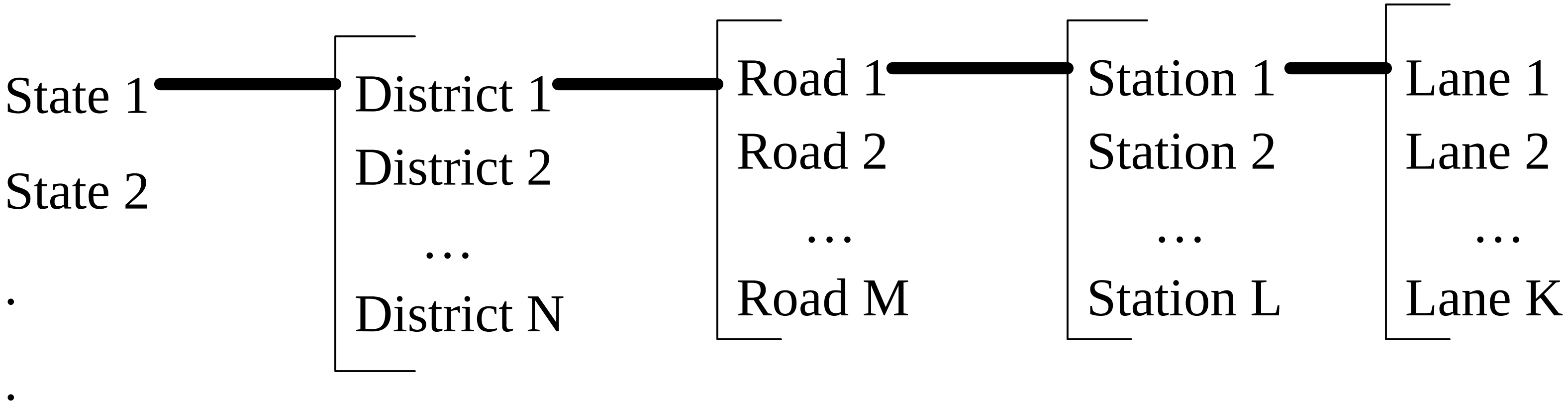
Large Scaled ITS Data Archiving

- Centralize it (It should never be tried by individuals using their desktop PCs.)
- Allocate financial, material, and human resources
- Start with a strong commitment: Archiving can never be stopped, if started.
- Analyze data characteristics and all logistics before archiving

Data Hierarchies

Spatial Hierarchy





State 1
State 2
.
.

District 1
District 2
…
District N

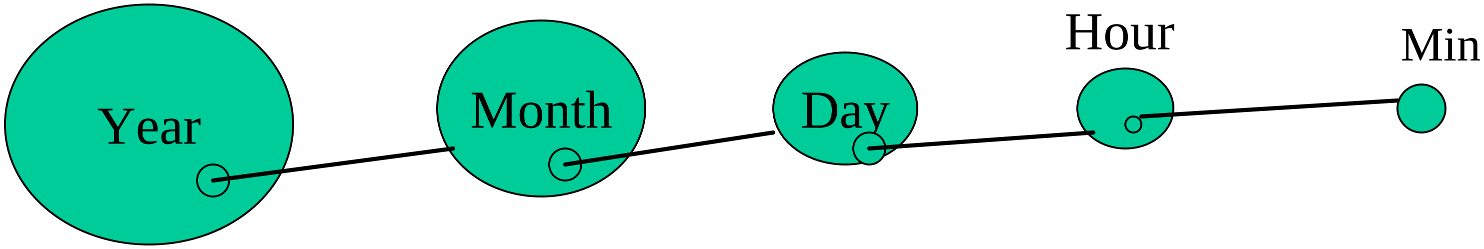
Road 1
Road 2
…
Road M

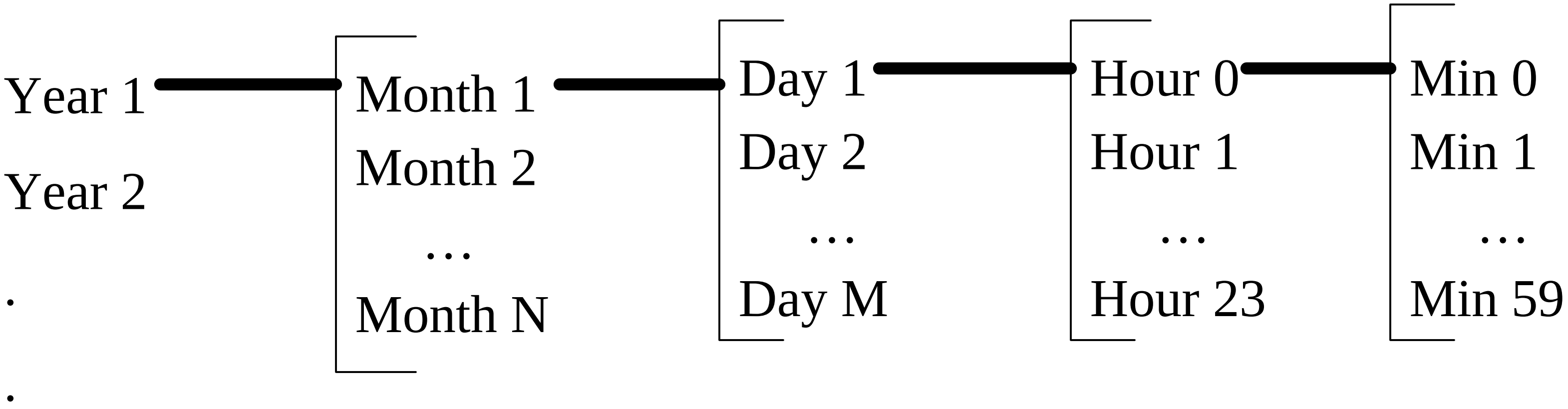
Station 1
Station 2
…
Station L

Lane 1
Lane 2
…
Lane K

Data Hierarchies, Cont

Temporal Hierarchy





Year 1
Year 2
.
.

Month 1
Month 2
…
Month N

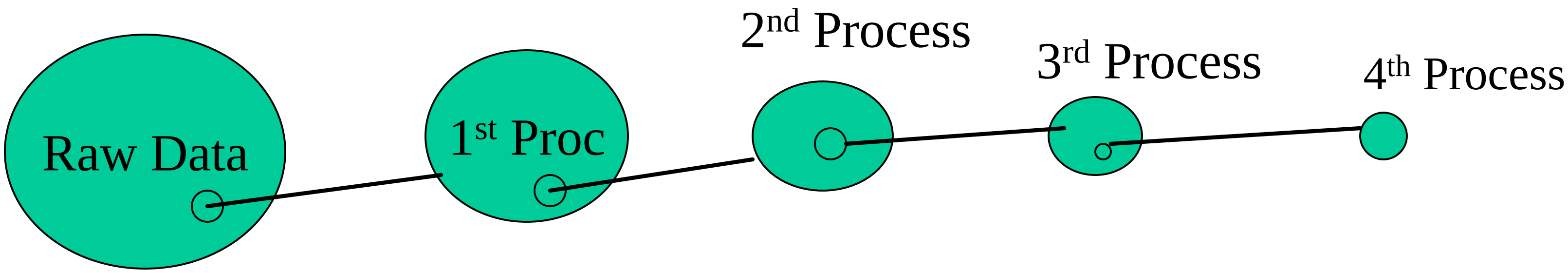
Day 1
Day 2
…
Day M

Hour 0
Hour 1
…
Hour 23

Min 0
Min 1
…
Min 59

Data Hierarchies, Cont

Computational Hierarchy



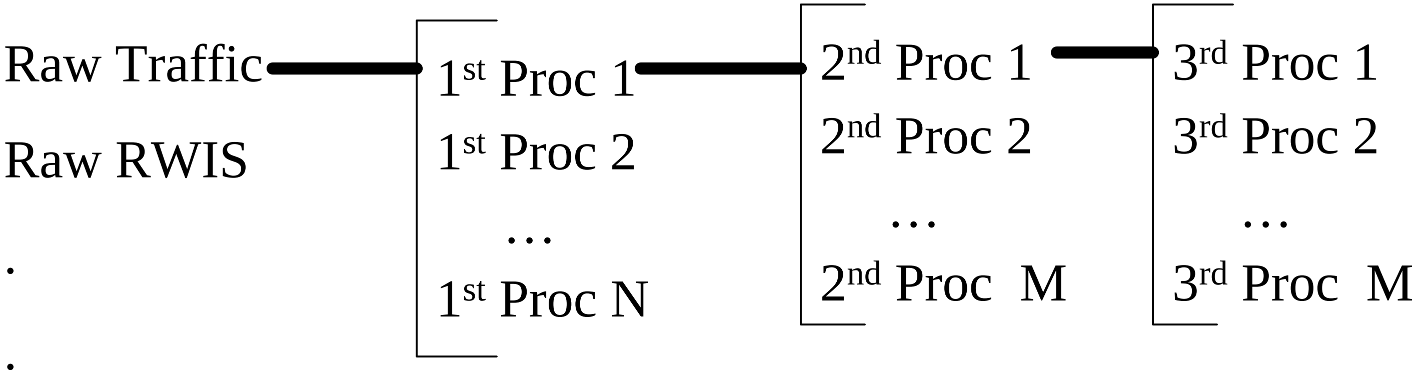
Raw Data

1st Proc

2nd Process

3rd Process

4th Process



Raw Traffic

Raw RWIS

.

.

1st Proc 1

1st Proc 2

…

1st Proc N

2nd Proc 1

2nd Proc 2

…

2nd Proc M

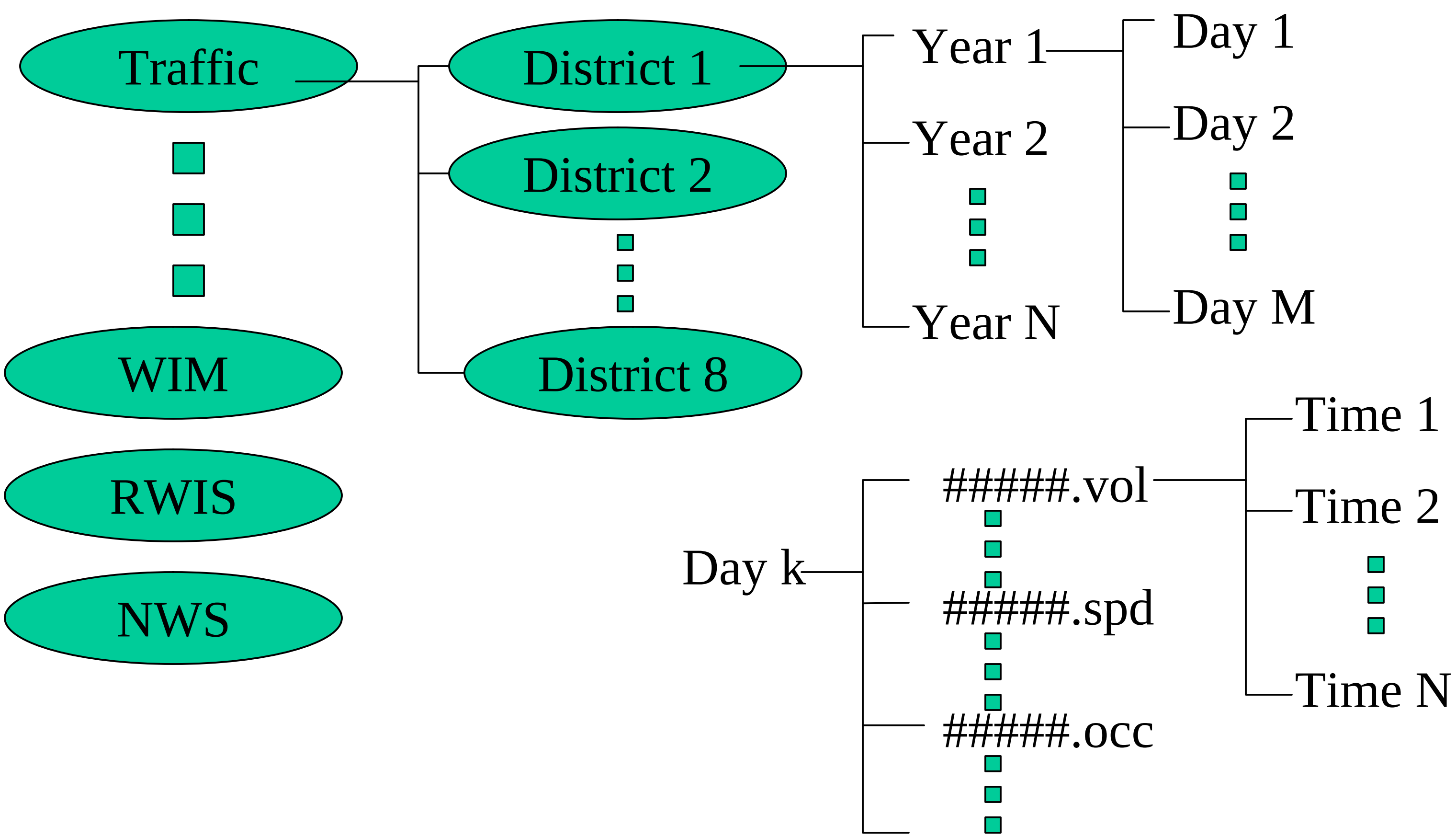
3rd Proc 1

3rd Proc 2

…

3rd Proc M

TDRL Archive Hierarchy Model



Raw Data Archiving

- Most important and critical step
- Collect data at the smallest time scale (highest sampling rate) as possible
- Try to achieve never fail robustness; real-time data can never be reproduced once it is lost. Called Write Once Read Many (WORM) data

Archiving Choices

- Binary (or Text) Zip Compressed Files
- CDF, HDF
- RDBMS

Desirable Properties of ITS Data Archive

- Small Size
- Fast Retrieval
- Portable Between Different OS
- Low Initial Investment and Maintenance Cost
- Metadata Capability
- Open Standard

Common Data Format (CDF)

- Developed for NASA Climate Data System at National Space Center Data Center
- Self-describing data abstraction for the storage and manipulation of multi-dimensional data (metadata)
- Transparent data format
- Transparent data compression
- Efficient sparse record handling
- Platform Independent
- API available in C, FORTRAN, Java, and Pearl

Conclusions

- Once archiving is started, it can never be stopped or failed. You must have a robust plan.
- Before starting archiving ITS data, every possible scenario must be reviewed
- Archiving data models need further study
- Large scaled data archiving is a challenging task requiring resources. It should not be considered as light tasks